In search of isoglosses: continuous and discrete language embeddings in Slavic historical phonology

Cathcart, Chundra ; Wandl, Florian

DOI: https://doi.org/10.18653/v1/2020.sigmorphon-1.28

In search of isoglosses: continuous and discrete language embeddings in Slavic historical phonology

Chundra A. Cathcart[1,2] **and Florian Wandl**[3]

[1]Department of Comparative Language Science, University of Zurich
[2]Center for the Interdisciplinary Study of Language Evolution, University of Zurich
[3]Slavisches Seminar, University of Zurich
{chundra.cathcart,florian.wandl}@uzh.ch

Abstract

This paper investigates the ability of neural network architectures to effectively learn diachronic phonological generalizations in a multilingual setting. We employ models using three different types of language embedding (dense, sigmoid, and straight-through). We find that the Straight-Through model outperforms the other two in terms of accuracy, but the Sigmoid model's language embeddings show the strongest agreement with the traditional subgrouping of the Slavic languages. We find that the Straight-Through model has learned coherent, semi-interpretable information about sound change, and outline directions for future research.

1 Introduction

Historical phonology is an important area of diachronic linguistics, allowing scholars to explore the space of possible sound change trajectories and resulting synchronic patterns, as well as posit degrees of relatedness between languages on the basis of sound changes shared across them. The latter practice traditionally involves the identification of innovations that are probative with respect to historical subgrouping. The internal genetic structure of many linguistic groups is uncontroversial. For others, scholars disagree in terms of which isoglosses are relevant to subgrouping, and whether the relevant features are indeed shared across groups of languages. The use of computational methods has aided in resolving a number of outstanding questions in diachronic linguistics, though little work has been done assessing the ability of computational models to learn meaningful patterns of sound change as well as capture language-level information that may bear on degrees of genetic relatedness.

This paper employs a neural encoder-decoder architecture to analyze patterns of sound change among Slavic languages, training a series of models on data from an etymological dictionary. Following the standard practice in multilingual NLP tasks, we make use of language embeddings concatenated to the model input. We make use of three different types of language embedding, comprising continuous real-valued DENSE, SIGMOID (defined on the $[0, 1]$ interval), and binary STRAIGHT-THROUGH embeddings. We assess the accuracy with which these encoder-decoder models predict held-out forms in contemporary Slavic languages from their corresponding Proto-Slavic input. We provide a detailed error analysis, observing differences across models in terms of the types of error introduced. We measure the extent to which the language embeddings learned by each model recapitulate the the most commonly accepted subgrouping of the Slavic languages. Finally, we assess the interpretability of the straight-through embedding, investigating the degree to which embeddings in binary latent space represent meaningful information regarding sound change.

We find that the model with straight-through language embeddings outperforms the Dense and Sigmoid models in terms of accuracy. At the same time, the language embeddings learned by the Sigmoid model display a signal that shows the highest agreement out of the three models with received wisdom regarding the dialect grouping of Slavic languages. We find that the latent binary representations learned capture meaningful and coherent information regarding sound patterns. We outline future directions for research using latent binary embeddings in neural historical phonology.

2 Background

The Slavic branch of Indo-European is traditionally divided into East, West, and South Slavic groups. Many of the oldest and most de-

Proceedings of the Seventeenth SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 233–244

https://doi.org/10.18653/v1/P17

cisive isoglosses differentiating the Slavic languages are phonological in nature (cf. Shevelov 1964, Carlton 1991). For instance, tautosyllabic Proto-Slavic vowel+liquid sequences were subject to METATHESIS or re-ordering in West and South Slavic languages, whereas East Slavic languages underwent PLEOPHONY, inserting a vowel between the liquid and the following consonant. Variation between liquid metathesis and pleophony, accompanied by language-specific vowel changes, can be seen in the cognates Russian *górod*, Ukrainian *hórod*, Croatian *grâd*, Czech *hrad* (< *gôrdŭ 'city'); the *h-* found in Ukrainian (East Slavic) and Czech (West Slavic) shows also that certain shared features do not cleanly follow the taxonomy defined above.

It is traditionally assumed that the tripartite classification of Slavic either reflects the dialectal diversity of the so-called Slavic homeland, most probably situated on the outskirts of the Carpathian Mountains, or emerged as a result of the great Slavic expansion in the 6th century AD (Bräuer, 1961; Hock, 1998). The extensive study of loanwords, however, suggests that post-expansional Slavic was, despite the vast territory it occupied, still uniform. There seem to have been no significant differences between Slavic spoken in areas located as far away from each other as the Baltic sea and the Peloponnese, at least with regard to phonology. It has therefore been argued that it is this post-expansional Slavic that constitutes the ancestor of all Slavic languages and not the Slavic language spoken in the homeland (Holzer, 1995). One of the arguments put forward in support of this claim is the still largely reconstructible post-Proto-Slavic dialect continuum (Holzer, 1997). One objective of this paper to assess the degree to which neural models recapitulate the uncontroversial subgrouping of Slavic as an indicator of whether they are capable of resolving outstanding issues in the field.

3 Related Work

A growing body of research assesses the information captured by language embeddings trained on large data sets using neural models. There is some debate as to whether embeddings learned in these tasks can pick up on genetic signal (Östling and Tiedemann, 2017; Tiedemann, 2018), or whether the information learned represents structural similarity (Bjerva et al., 2019). The majority of work of language embeddings involves models trained on large parallel corpora. Meloni et al. (2019) approach the issue of sound change using a GRU-based neural machine translation model with soft attention to reconstruct Latin forms from contemporary Romance reflexes; the authors employ language embeddings, but do not provide an analysis of the information captured by these embeddings. Phylogenetic approaches to sound change and the reconstruction of word forms incorporate a highly articulated genetic representation of language relatedness (Hruschka et al., 2013; Bouchard-Côté et al., 2013), but employ simplified representations of sound change in comparison to what can be captured by recurrent neural networks; at the same time, phylogenetic work explicitly models intermediate stages of change, a potential challenge for RNNs, which are better suited to learning patterns resulting from the telescoping of multiple changes. Related work seeks to disentangle genetic and areal pressures in shaping cross-linguistic patterns (Daumé III, 2009; Murawaki and Yamauchi, 2018; Cathcart, 2019, 2020b,a).

In general, while the signal learned by embeddings can be analyzed via visualization techniques (Maaten and Hinton, 2008), it is a challenge to link the behavior of embeddings to individual features in the data analyzed. This difficulty undoubtedly stems in part from the fact that embeddings are generally continuous, lacking the sparsity or discreteness needed to identify the behavior of the neural model when features are active or inactive. This issue has been addressed by the development of de-noising approaches designed to induce sparsity (Subramanian et al., 2018).

Binary latent variables are of key interest to linguistic questions, but pose many challenges for inference. Binary latent variable models such as the Indian Buffet Process (IBP, Ghahramani and Griffiths, 2006) have been used in some applications in computational phonology and typology (Doyle et al., 2014; Murawaki, 2017) using a combination of Gibbs Sampling and updates from the Metropolis-Hastings algorithm or Hamiltonian Monte Carlo, but it is not clear that these inference procedures are scalable to neural models. Discrete variables pose problems for differentiability in gradient-based optimization algorithms; marginalizing out all possible combinations of binary variables is generally unfeasible for binary latent variables. Variational approaches

have attempted to circumvent this issue via the concrete (alternatively, Gumbel-Softmax) distribution (Maddison et al., 2017; Jang et al., 2017), which extends the reparameterization trick to categorical distributions and which produces gradient estimates that have lower variance than standard estimation techniques (Williams, 1992) but are still biased; subsequent approaches reduce bias but are less straightforward to implement (Grathwohl et al., 2018; Liu et al., 2019).

While concrete-distributed versions of the IBP have been used in neural models (Singh et al., 2017; Kessler et al., 2019), this work is limited to variational autoencoders, which use amortized variational inference to learn latent representations from the data via a global inference network; encoder-decoder mechanisms with attention like the one used in this paper cannot exploit this property of the data; training the latent variable with stochastic variational inference, while theoretically possible, is considerably more difficult (Kim et al., 2018). As an alternative, we use straight-through (ST) embeddings (Bengio et al., 2013; Courbariaux et al., 2016) in a maximum likelihood framework. Straight-through layers are discrete but have underlying continuous weights; model output is predicted on the basis of the discrete representation, while model loss is differentiated with respect to the continuous underlying weights. While this approach has the same problems with biased estimates as the concrete distribution, it is straightforward to implement. We compare the quality of straight-through embeddings to embeddings with no activation and embeddings with sigmoid activation.

4 Data

Our data set consists of Proto-Slavic etyma and corresponding reflexes in medieval and modern Slavic languages taken from a digitized version of a Slavic etymological dictionary (Derksen 2007; for alternative reconstructions see Holzer 1995; Andersen 1998). In order to minimize the chance of introducing morphologically non-congruent forms into our data set, we extracted the first form provided for each Slavic language in each entry, since these are most likely to agree morphologically with the Proto-Slavic headword.

We converted forms in modern Slavic languages to a narrow phonetic representation using IPA transcriptions from Wiktionary (`https://www.wiktionary.org`), which were used to train a neural encoder-decoder; these models were used to obtain IPA transcriptions for forms not in Wiktionary, and a portion was checked manually. In several cases we reconciled sources used in the etymological dictionary (e.g., Pleteršnik, 1894) with contemporary standardized orthographies, and made use of phonetic descriptions for languages where the training data were problematic (Schuster-Šewc, 1968; Lencek, 1982; Scatton, 1984; Comrie and Corbett, 1993; Ternes and Vladimirova-Buhtz, 1990; Landau et al., 1995; Šuštaršič et al., 1995; Dankovičová, 1997; Jassem, 2003; Gussmann, 2007; Stadnik-Holzer, 2009; Hanulíková and Hamann, 2010; Mojsijenko et al., 2010; Yanushevskaya and Bunčić, 2015; Howson, 2017, 2018; Pompino-Marschall et al., 2017). For the medieval languages Old Church Slavic and Church Slavic, orthographic forms were converted to a broad phonemic transcription based on Lunt (2001). Suprasegmental features were marked for all modern languages (pitch accent for Slovene and BCS and primary stress for the remainder; for consistency, we chose to mark primary stress on monosyllables in stress-timed languages). We excluded languages with fewer than 100 forms in the etymological dictionary (this resulted in the omission of Macedonian, Polabian and Slovincian).

We took additional steps to remove morphological mismatches in the data set. For Bulgarian verbs, which reflect the Proto-Slavic 1sg present in their citation form, we replaced the Proto-Slavic headword (the infinitive form by default) with a morphologically congruent form, and excluded a small number of forms based on athematic verbs. Additionally, Proto-Slavic adjectives are always given in the nominal or short form, although contemporary Slavic languages often reflect the so-called long form, which arose from the addition of an inflected element *-jь̌ to the ending; we converted short Proto-Slavic adjectives to their long form in the appropriate contexts. We tried to ensure that Proto-Slavic verbs matched their reflexes according to the presence/absence of reflexive morphology and preverbs. Additionally, the original data source contains multiple gender inflections for certain Proto-Slavic etyma (e.g., *àblъ̌ko n., *àblъ̌ka f., and *àblъ̌kъ̌ m. for ‘apple’), which are linked to the same reflexes irrespective of the reflexes’ gender; for such forms, we discarded etymon-reflex pairs with mismatched

Language	Glottocode	# reflexes
Russian (Rus)	russ1263	1572
Slovene (Sln)	slov1268	1462
Serbo-Croatian (BCS[M])	sout1528	1434
Czech (Cze)	czec1258	1377
Polish (Pol)	poli1260	1282
Slovak (Slk)	slov1269	1091
Old Church Slavic (OCS)	chur1257	1097
Bulgarian (Bul)	bulg1262	950
Church Slavic (CS)	chur1257	392
Ukrainian (Ukr)	ukra1253	301
Upper Sorbian (USo)	uppe1395	243
Lower Sorbian (LSo)	lowe1385	120
Belarusian (Bel)	bela1254	79
Total		11400

Table 1: Number of forms in each language in data set, along with closest matching glottocodes.

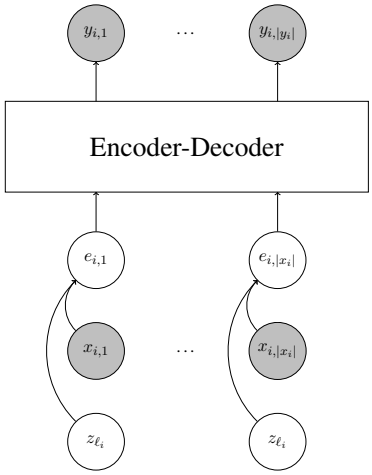

Figure 1: Basic schema of architecture used in this paper; for each input-output pair $\mathbf{x}_i, \mathbf{y}_i$, an embedding associated with the language ID for index i is concatenated to a one-hot representation of the input.

gender. Ultimately, this process yielded 11400 forms in 13 languages (see Table 1), and allowed us to rid the data set of a large number (albeit not the entirety) of morphological mismatches.

5 Method

To learn mappings between Proto-Slavic etyma and the Slavic reflexes that descend from them, we use an LSTM Encoder-Decoder with 0th-order hard monotonic attention (Wu and Cotterell, 2019), trained on all languages in our data set. The basic model architecture used for the experiments in this study has the following structure (schematized in Figure 1): a trainable language-level embedding is concatenated to a one-hot representation of each input segment at each input time step; each concatenation is fed to a Dense layer (with no activation) to generate a embedding for each time step that encodes information about the input phoneme and language ID of the reflex; these embeddings subsequently are fed to the encoder-decoder in order to generate the output. The parameters of the encoder-decoder architecture are shared across languages in the data set; the sole language-specific variable employed is the language-level embedding fed to the model.

In all experiments, we set the dimension of the language-level embedding and the language/character embedding to 128, and the hidden layer dimension to 256. In our experiments, we employ different representations of the language-level embedding, including a dense layer with no activation (DENSE model), a dense layer with sigmoid activation, (SIGMOID model) and a dense layer with straight-through activation (ST model), which uses the Heaviside step function (negative values map to 0, non-negative values to 1). We train our model for 200 epochs with a batch size of 256 using the Adam optimizer with a learning rate of .001 with the objective of minimizing the mean categorical cross-entropy between the predicted and observed distributions of the output. To evaluate model performance, we carry out K-fold cross-validation ($K = 10$), randomly holding out 10% of the forms in each language, and greedily decoding the held-out forms using the trained model. For additional analyses regarding the interpretability of the embeddings learned, we train the model on all forms in the data set. Models are implemented in Keras (Chollet, 2015) and Larq (Geiger and Team, 2020).[1]

6 Results

6.1 Accuracy

We assess the accuracy of each model by generating held-out forms on the basis of the language ID of the form and the Proto-Slavic etymon from which the form descends, greedily decoding on the basis of the trained model. We measure accuracy in terms of word error rate (WER), which gives the proportion of incorrectly generated forms, and the phoneme error rate (PER), which we define as the Levenshtein edit distance between generated and ground truth strings divided by the length of the longer form. Accuracy measures are found in Table 2. The ST model shows the best performance,

[1]Code accompanying this paper is available at https://github.com/chundrac/slav-dial/tree/master/SIGMORPHON_2020

	WER	PER
Dense	0.535	0.143
Sigmoid	0.559	0.151
ST	**0.530**	**0.140**

Table 2: Mean word error and phoneme error rate for each model

followed by the Dense and Sigmoid models. Figure 2 shows WER and PER for each model plotted by the log number of training examples for each language in our data set. There is at best a very weak negative correlation between error rates and training example frequencies; the worst performance seems to be restricted to four languages (Belarusian, Lower Sorbian, Ukrainian, and Upper Sorbian), which vary in training data frequency, but in our impression posed the most difficulties for phonetic conversion. Old Church Slavic and Church Slavic show the highest accuracy; forms in these languages tend to be close to their Proto-Slavic ancestral forms,[2] and were straightforward to convert to IPA.

6.2 Error analysis

6.2.1 Quantitative error analysis

We wish to obtain a fuller picture of the errors made by our models, and in particular, whether different models produce different types of errors. We analyze errors according to a taxonomy inspired by Gorman et al. (2019). At a high level, errors can be divided according to whether they stem from mistakes in the data or are a result of model idiosyncrasies. Errors in the data (target errors) largely consist of morphologically non-congruent etymon reflex pairs that we were unable to detect *a priori*: for instance, the Slavic etymon *dĭlti 'to hollow, chisel' is paired with reflexes such as Czech *dlbsti*, which contains the cluster *-bs-* due to analogical influence; similarly, the etymon *majati 'wave, beckon' (inf.) is paired with OCS *namaiaaxǫ* (3pl impf.). Additionally, there exists the possibility of doublet reflexes in contemporary Slavic languages due to dialect borrowing (free variation errors), e.g., Russian *óblako* from Church Slavic (Vasmer, 1953-1958). Incorrect phonetic conversion is another source of errors of this type.

[2]Note that according to the common practice in etymological dictionaries, OCS snd CS forms are given in a normalized form not reflecting regional differences.

In terms of linguistic errors that are not direct artifacts of our data set, we are interested in the degree to which the models' behavior results in a specific set of error pattern types. We wish to measure the extent to which models introduce errors when decoding forms in a given language due to overgeneralization on the basis of forms seen in the training data for the SAME LANGUAGE. For instance, all models fail to learn the Upper Sorbian development *pr > [pʃ], erroneously generalizing the change *r > [ʀ] to an incorrect environment (e.g., PSl *pręsti 'spin' > [ˈpʀʲast͡ʃ], expected [ˈpʃast͡ʃ]). Additionally, because our model leverages global information shared across languages along with language-specific information, errors in one language involving the application of a sound law from a DIFFERENT SLAVIC LANGUAGE are a potential concern. For example, the Sigmoid model generates the erroneous BCS reflex [lĕt͡ɕæti] 'to fly' (< PSl *letěti, expected [lětjeti]); [æ] is attested only in OCS, CS, Russian, and Slovak. Of additional interest are errors where the model produces a rule that is unattested across the data set, and hence UNMOTIVATED by the data. For instance, the Sigmoid model generates BCS [ppĕːta] (< PSl *pętà 'heel', expected [pĕːta]); word-initial [pp-] is unattested in our data set, and the origin of this error is unclear.

We quantitatively assess the issues enumerated above in the manner described below. To assess the prevalence of target errors, we measure the extent to which models agree in terms of the data points for which poor performance is exhibited. We take this agreement as a proxy for errors in the data; if the same data points cause problems across models, this poor performance may be an artifact of morphological mismatches in the data or fewer examples in the training data than needed to learn the patterns for the data points in question. The agreement matrix in Table 3 shows that agreement levels are quite high, indicating that some errors may be due to artifacts of the data used.

To gain an overview of the error types made by the model, we use the attention mechanism of the trained models to obtain alignments between all Proto-Slavic etyma and attested reflexes as well as between Proto-Slavic etyma and erroneously produced reflexes. We extract sound changes operating between Proto-Slavic and daughter languages from these alignments (e.g., PSl *o > Slovak ɔ), which indicate whether a given edit is attested in

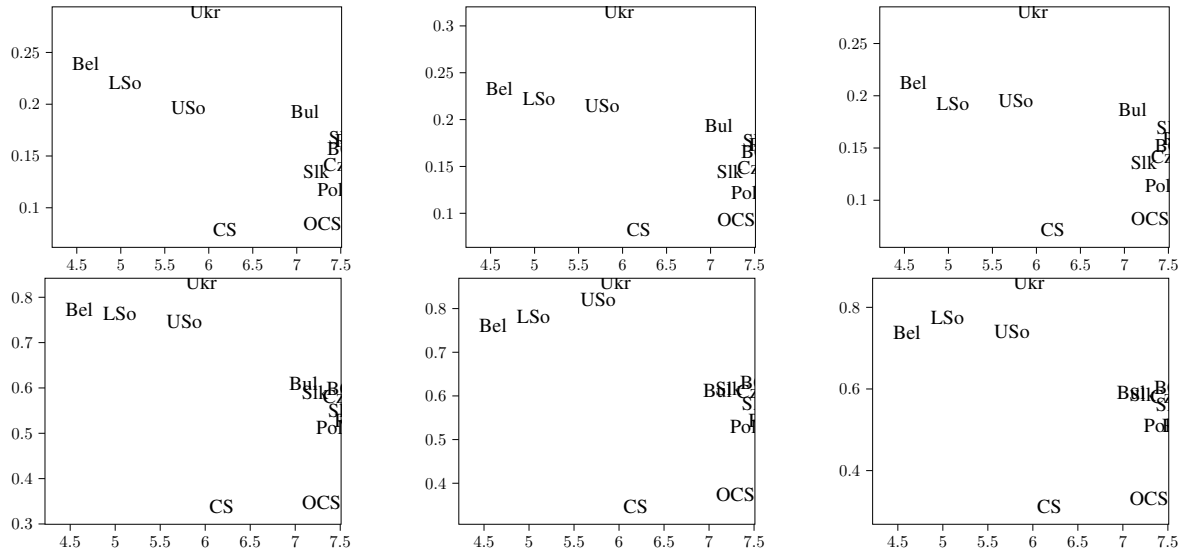

Figure 2: PER (top) and WER (top) values (y axis) plotted by the log number of training examples for each language (x axis), for Dense, Sigmoid and ST models (left to right)

Model	Dense	Sigmoid	ST
Dense	—	0.789	0.812
Sigmoid	0.824	—	0.827
ST	0.803	0.783	—

Table 3: Proportion of word errors produced by each model (rows) shared with other models (columns)

Model	SL	OL	U
Dense	0.551	0.105	0.342
Sigmoid	0.593	0.101	0.305
ST	0.617	0.101	0.281

Table 4: Proportion of errors produced by models that are present in the same language (SL), other languages (OL), or are unmotivated (U)

a language (irrespective of conditioning environment). We automatically annotate each erroneous edit according to whether it is attested in the same language as the decoded form in which it occurs (same language), if not, whether it is attested in another Slavic language (other language), or finally, if it is not attested in any Slavic language (unmotivated). Table 4 shows proportions of these error types produced by each model; the Sigmoid and ST models produce more other-language and unmotivated errors than the Dense model.

6.2.2 Qualitative error analysis

We present results of a detailed error analysis involving 422 forms spanning all languages in the data set where at least one of the three models produced an error. Roughly 15% of the forms surveyed contain some sort of morphological mismatch; many of these are trivial one-off analogical idiosyncrasies. In some cases, loanwords unmarked in the dictionary can be detected (cf. the example of Russian *óblako* mentioned above).

Annotated error types that occur more than once across all models include incorrect accent type (Dense: 18, Sigmoid: 12, ST: 10), accent misplacement (Dense: 40, Sigmoid: 40, ST: 32), consonant mismatches (Dense: 139, Sigmoid: 161, ST: 149), vowel quality mismatches (Dense: 192, Sigmoid: 219, ST: 195), vowel length mismatches (Dense: 35, Sigmoid: 47, ST: 31), and general segmental mismatches involving the erroneous substitution of a vowel for a consonant, or vice versa (Dense: 85, Sigmoid: 81, ST: 68). The ST model's overall higher performance bears out the larger-scale analysis of errors presented in the previous section.

Our manual error analysis was carried out by a single specialist; future research will involve more detailed error analyses carried out by multiple specialists in order to gauge inter-annotator reliability.

6.3 Genetic signal in embeddings

We wish to measure the degree to which the language-level embeddings learned by each model reflect received wisdom regarding the dialectal makeup of the Slavic languages. As stated previ-

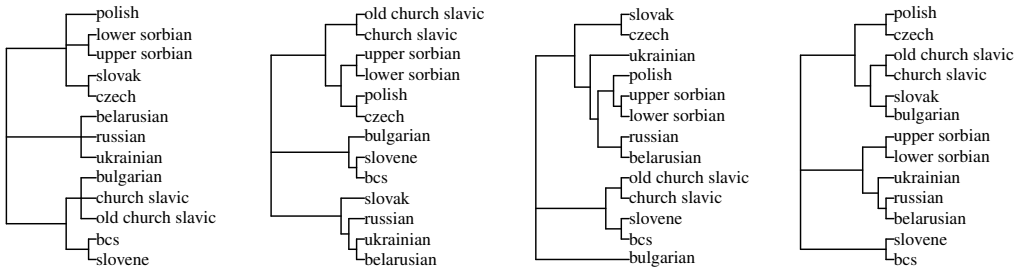

Figure 3: Reference phylogeny of Slavic languages and neighbor-joined trees from embeddings for Dense, Sigmoid, and Straight-Through models (left to right)

ously, languages employed in this paper are traditionally divided among East, South and West Slavic groups. To assess the signal contained by the embeddings, we generated trees from cosine distances between pairs of language embeddings learned by each model using neighbor joining (NJ, Saitou and Nei, 1987) as implemented in the R package `ape` (Paradis et al., 2019). These trees can be found in Figure 3 alongside a reference topology from Glottolog (Hammarström et al., 2017). The Sigmoid model's embeddings show the highest agreement with the Glottolog tree; the main discrepancies found are the placement of Bulgarian outside of South Slavic, as well as the placement of the Lechitic languages Polish, Upper Sorbian and Lower Sorbian within East Slavic. The ST embeddings show mixed performance; certain West and South Slavic languages are grouped correctly, but a large number of taxa are misplaced. We used the R package `Quartet` (Smith, 2019) to measure the generalized quartet distance (Pompei et al., 2011) between the reference tree and the trees constructed from the embeddings, equal to the number of four-taxon groups resolved differently across the two trees, divided by the number of resolved four-taxon groups found in the reference tree; lower values indicate greater agreement (Dense: 0.322, Sigmoid: **0.247**, ST: 0.368). It is possible that the ST model shows low agreement with the reference phylogeny but high accuracy because it has succeeded in detecting areal features that conflict with the traditional tripartite subgrouping. Further investigation into the treelikeness of each network (Wichmann et al., 2011) is needed in order to properly address this issue.

6.4 Interpretation of embeddings

A common goal of neural modeling with discrete latent variables is to learn sparse interpretable features. Ideally, activating or deactivating a single binary latent variable should correlate with the presence or absence of a meaningful feature in the model's output. Inducing the level of sparsity needed to generate such latent variables is an ongoing issue in the deep learning literature (Singh et al., 2017). Our models have not learned meaningful features in the sense that turning a single variable "on" or "off" can produce a meaningful feature of Slavic dialectology (e.g., the presence of liquid metathesis/pleophony); these processes appear to be distributed across multiple latent binary variables.

Figure 4: Active dimensions (white cells) in ST language embeddings

As shown in Figure 4, language-level straight-through embeddings are far from sparse; of the 128 embedding dimensions, only 1 is inactive across languages. Individual language embeddings contain between 32 and 61 active dimensions. Preliminary attempts to turn individual dimensions "on" and feed the latent representation to the encoder-decoder along with a Proto-Slavic input do not produce interpretable or coherent results; it appears to be the case that it is not individual dimensions, but interactions between them, that influence the behavior of the decoder.

6.4.1 Nearest neighbors

Feeding all possible 2^{128} combinations of embedding values to the model is computationally infeasible, though it might allow us to discover which feature combinations are responsible for certain types of behavior of the encoder-decoder. In order to gain a better understanding of the behavior of these dimensions individually and as a group, we explore the NEAREST NEIGHBORS in embed-

ding space of reflexes for languages in our data set by altering the values of each variable in our language embeddings, and feeding these altered embeddings to the model architecture along with a Proto-Slavic etymon and observing the set of resulting outputs. Specifically, we take language-level embeddings and alter each of the embedding's 128 dimensions.

By feeding these nearest neighbors to the encoder-decoder along with a Proto-Slavic etymon, it is possible to see how perturbations of an embedding result in different outputs from the expected contemporary Slavic reflex. In general, different perturbations often result in the same output form, indicating that the embedding space is perhaps less sparse than necessary, and more compact representations can be learned without losing information. To give a concrete example of this phenomenon, the nearest neighbors of Polish [ˈmɨdwɔ] 'soap, lather' (< PSl *mỳdlo) in embedding space yield only thirteen unique forms ([ˈmiloˌ], [ˈmidlɔ], [mĭːdlɔ], [ˈmɨdwɔ], [mɨːlo], [ˈmɨdlɔ], [mɪllɐ], [mɪllɔ], [mĭlɔ], [mɪlɔ], [ˈmilɔ], [mĭːlɔ]); interestingly, there is no evidence for the otherwise naturalistic and plausible sound change *dl > [ll] in our data set.

Based on a qualitative appraisal of these nearest neighbors, it does not appear to be the case that the ST model has learned to entirely disentangle orthogonal developments in historical phonology. The unique nearest neighbors of Russian [məlʲˈko] 'milk' (< PSl *melkò), [mlɛkɔ], [ˈmlɛko], [mlĕːkɔ], [mlɪˈko], [mlĕːko], [ˈmlɛkɔ], [məlʲˈko] and [mlɛˈkɔ], appear to show that our model learns patterns of pleophony/liquid metathesis and vowel change jointly, rather than learning disentangled abstractions (though interestingly, the same word in Polish has the neighbors [ˈmʲɪlkɔ] and [mlʲɪˈkɔ], showing metathesis independent of vowel quality). It is not clear, however, that this behavior goes against the received wisdom of Slavic linguistics; the operation of liquid metathesis or pleophony among Slavic languages is generally thought to be a change that has an early common origin but developed in different dialect-specific directions (Shevelov, 1964). Ultimately, this architecture shows the potential to generate typologically meaningful (i.e., naturalistic) but also novel representations of hypothetical Slavic reflexes.

6.4.2 Sampling from the latent space

An issue that arises in the use of latent variable models, particularly in the context of linguistic typology, concerns the coherence of the representations that they learn. If we randomly traverse our models' latent variable space or interpolate between representations, how likely are we to encounter a plausible unattested sister language of the languages attested in our data set? We briefly explore this question by randomly sampling 100 embeddings from variously parameterized distributions and feeding them to our models, along with a set of 100 randomly chosen Proto-Slavic etyma. For each etymon, we feed zero-mean Gaussian samples with standard deviation $\sigma \in \{.01, .1, 1, 10\}$ to the Dense model; symmetric Beta samples with shape parameters $\alpha = \beta \in \{.01, .1, 1, 10\}$ to the Sigmoid model; and Binomial samples with probability $p \in \{.2, .4, .6, .8\}$ to the ST model (all samples have the same dimension as our learned embeddings). Qualitatively speaking, output forms randomly generated by the ST model are consistently well formed and coherent across parameterization regimes. Conversely, when σ is greater than .01 (roughly equivalent to the empirical standard deviation of the learned embeddings), the Dense model often generates unrealistic strings (e.g., [bɬɬˈːɬɬə]), and when σ is very small, forms are coherent but there is virtually no variation; for the Sigmoid model, the strings become more realistic looking as $\alpha = \beta$ increases (the majority of values for the learned Sigmoid embeddings are close to .5). To highlight a related discrepancy, we observe the average number of unique outputs generated by each regime in each model (Dense: 3.21, 24.02, 93.08, 61.3; Sigmoid: 96.23, 94.14, 67.39 22.74; ST: 21.3, 24.3, 24.06, 20.1); the quantity of unique outputs stays constant across all regimes for the ST model, along with their quality.

Additionally, we wish to explore the extent to which samples from latent variable space generate realistic sound changes and plausible sound patterns. While certain diachronic trajectories can lead to the emergence of "crazy" rules (Bach and Harms, 1972; Buckley, 2000) and unnatural phonotactic restrictions (Beguš and Nazarov, 2017), we might expect the relatively infrequent nature of these phenomena to somehow be captured by the behavior of models like the ST model. To address this question, we feed sets

of hypothetical well-formed Proto-Slavic phonological neighbors (generated by taking 12 etyma from our data set and generating echo-forms to create a cohort of forms differing according to initial *p-/t-/k-/b-/d-/g-; we exclude hypothetical forms with velar-front vowel sequences, which would have been affected by palatalization) to the ST model, randomly sampling binary latent embeddings from the binomial distribution with probabilities $\{.2, .4, .6, , 8\}$. For each probability regime, we attempt to evaulate the relative frequency of unnatural sound patterns displayed by these hypothetical forms' descendants; if our model embodies not only plausible but probable behavior, we predict that these etymological phonological neighbors, which differ only according to the word-initial consonant, should frequently yield similar echo-forms, and that other patterns may arise less frequently. For each pair of outputs within each cohort (with stress marking removed), we divide the number of agreeing final segments by the mean of the two strings' lengths, and report the proportion of pairs for which this value is greater than .5 (indicating greater agreement); these values are 0.427, 0.546, 0.574, and 0.526 for each respective probability regime, indicating that generated outputs tend not to be very echo-like. To exemplify, a representative sample output for *{p,t,b,d}ĭrtĭ comprises the forms [ˈpr̩t͡ɕ], [ˈtr̩t], [ˈbr̩t͡s], [ˈdr̩t]. While long-distance assimilatory and dissimilatory processes operating between the left and right word edge are not unknown cross-linguistically, we believe that changes where differences in word-initial segments trigger divergent word-final reflexes should be rare, rather than typical. Further refinement of metrics designed to assess the validity of output patterns is much needed.[3]

From this small and rather premature investigation, it appears that the latent variable design space represented by the ST model generates coherent, realistic-looking output, but the frequency distributions of patterns in its output may not reflect cross-linguistic frequency distributions. A more in-depth analysis along these lines is outside the scope of this paper, and methods seeking to derive typological generalizations should include data from multiple families; at the same time, the issues raised here potentially bear on our understanding of the diachronic basis of synchronic patterns in phonology.

7 Discussion and outlook

This paper investigated the performance of multiple neural models in capturing patterns of sound change across Slavic languages. We found that a model with binarized straight-through language-level embeddings outperformed other models in terms of accuracy, and shows great potential for learning coherent and interpretable information regarding sound change. We found that the discrete features learned by our model appear for the most part to correspond to meaningful, realistic variation in sound patterns, though representations are not particularly sparse. Additionally, randomly sampling from discrete latent space tended to consistently generate coherent output; the preliminary attempts that we made to assess the likelihood of observing these samples in naturalistic contexts can be expanded considerably.

We used straight-through embeddings as a low-cost alternative to more involved means of training discrete latent variables. In the immediate future, we plan to extend our approach to make use of variational approaches, the flexibility of which may help in inducing sparsity in order to learn more meaningful, realistic representations (there is additionally room for exploration of simpler approaches that we did not make use of in this paper, such as dropout regularization); however, since our encoder-decoder model is different from the autoencoding models used in previous work, directly extending these methods presents a challenge that require considerable experimentation to overcome (an early attempt to adopt the IBP prior of Singh et al. 2017 was unsuccessful, as the monotonically decreasing prior probabilities rarely yielded non-zero values; thus far, attempts to weight the KL divergence term have not yielded success). Nevertheless, as low-variance, low-bias techniques for inferring discrete variables in neural models progress, we believe that they will be an increasingly valuable means of capturing meaningful, interpretable features in multilingual neural tasks like this paper's.

[3] Indeed, taking the proportion of agreeing final segments as a measure of naturalness would classify changes resulting from certain types of tonogenesis to be unnatural, e.g., *pa, *ba > Vietnamese *pa*, *pà* (Haudricourt, 1954), since dissimilarity in initial consonants often leads to dissimilarity at the right word edge.

References

Henning Andersen. 1998. Slavic. In Paolo Ramat and Anna G. Ramat, editors, *The Indo-European languages*, Routledge language family descriptions, page 415–453. Routledge, London and New York.

Emmon Bach and Robert T. Harms. 1972. How do languages get crazy rules? In *Linguistic Change and Generative Theory*, pages 1–21, Bloomington.

Gašper Beguš and Aleksei Nazarov. 2017. Lexicon against naturalness: Unnatural gradient phonotactic restrictions and their origins.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

Johannes Bjerva, Robert Östling, Maria H. Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. What do language representations really represent? *Computational Linguistics*, 45(2):381–389.

Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110:4224–4229.

Herbert Bräuer. 1961. *Slavische Sprachwissenschaft I. Einleitung, Lautlehre*. Walter de Gruyter Co., Berlin.

Eugene Buckley. 2000. On the naturalness of unnatural rules. In *Proceedings from the Second Workshop on American Indigenous Languages. UCSB working papers in linguistics*, volume 9, pages 1–14.

Terence R. Carlton. 1991. *Introduction to the phonological history of the Slavic languages*. Slavica Publishers, Columbus, Ohio.

Chundra A. Cathcart. 2019. Toward a deep dialectological representation of Indo-Aryan. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 110–119.

Chundra A. Cathcart. 2020a. Dialectal layers in West Iranian: a Hierarchical Dirichlet Process approach to linguistic relationships. *arXiv preprint arXiv:2001.05297*.

Chundra A. Cathcart. 2020b. A probabilistic assessment of the Indo-Aryan Inner-Outer hypothesis. *Journal of Historical Linguistics*, 10:42–86.

François Chollet. 2015. Keras. https://github.com/fchollet/keras.

Bernard Comrie and Greville G. Corbett. 1993. *The Slavonic languages*. Routledge reference. Routledge, London.

Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. 2016. Binarized neural networks: Training deep neural networks with weights and activations constrained to $+1$ or -1. *arXiv preprint arXiv:1602.02830*.

Jana Dankovičová. 1997. Czech. *Journal of the International Phonetic Association*, 27(1-2):77–80.

Hal Daumé III. 2009. Non-parametric Bayesian areal linguistics. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, pages 593–601, Boulder, CO. Association for Computational Linguistics.

Rick Derksen. 2007. *Etymological dictionary of the Slavic inherited lexicon.* Online version (https://ordbog.oesteuropastudier.dk/) accessed 1 February 2020. Brill, Leiden.

Gabriel Doyle, Klinton Bicknell, and Roger Levy. 2014. Nonparametric learning of phonological constraints in optimality theory. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1094–1103.

Lukas Geiger and Plumerai Team. 2020. Larq: An open-source library for training binarized neural networks. *Journal of Open Source Software*, 5(45):1746.

Zoubin Ghahramani and Thomas L Griffiths. 2006. Infinite latent feature models and the indian buffet process. In *Advances in neural information processing systems*, pages 475–482.

Kyle Gorman, Arya D. McCarthy, Ryan Cotterell, Ekaterina Vylomova, Miikka Silfverberg, and Magdalena Markowska. 2019. Weird inflects but ok: Making sense of morphological generation errors. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 140–151.

Will Grathwohl, Dami Choi, Yuhuai Wu, Geoffrey Roeder, and David Duvenaud. 2018. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. *ICLR*.

Edmund Gussmann. 2007. *The phonology of Polish*. The phonology of the world's languages. Oxford University Press, New York.

Harald Hammarström, Robert Forkel, and Martin Haspelmath. 2017. Glottolog 3.3. Max Planck Institute for the Science of Human History.

Adriana Hanulíková and Silke Hamann. 2010. Slovak. *Journal of the International Phonetic Association*, 40(3):373–378.

André-Georges Haudricourt. 1954. De l'origine des tons en vietnamien. *Journal asiatique*, 242:69–82.

Wolfgang Hock. 1998. Das Urslavische. In Peter Rehder, editor, *Einfürung in die slavischen Sprachen*, pages 17–34. Wissenschaftliche Buchgesellschaft, Darmstadt.

Georg Holzer. 1995. Die Einheitlichkeit des Slavischen um 600 n. Chr. und ihr Zerfall. *Wiener Slavistisches Jahrbuch*, 41:55–89.

Georg Holzer. 1997. Zum gemeinslavischen Dialektkontinuum. *Wiener Slavistisches Jahrbuch*, 43:87–102.

Phil Howson. 2017. Upper Sorbian. *Journal of the International Phonetic Association*, 47(3):359–367.

Phil Howson. 2018. Upper Sorbian – CORRIGENDUM. *Journal of the International Phonetic Association*, 48(3):389–389.

Daniel J. Hruschka, Simon Branford, Eric D. Smith, Jon F. Wilkins, Andrew Meade, Mark Pagel, and Tanmoy Bhattacharya. 2013. Detecting regular sound changes in linguistics as events of concerted evolution. *Current Biology*, 25:1–9.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. *ICLR*.

Wiktor Jassem. 2003. Polish. *Journal of the International Phonetic Association*, 33(1):103–107.

Samuel Kessler, Vu Nguyen, Stefan Zohren, and Stephen Roberts. 2019. Indian buffet neural networks for continual learning. *4th workshop on Bayesian Deep Learning (NeurIPS 2019)*.

Yoon Kim, Sam Wiseman, and Alexander M Rush. 2018. A tutorial on deep latent variable models of natural language. *arXiv preprint arXiv:1812.06834*.

Ernestina Landau, Mijo Lončarić, Damir Horga, and Ivo Škarić. 1995. Croatian. *Journal of the International Phonetic Association*, 25(2):83–86.

Rado Ludovik Lencek. 1982. *The structure and history of the Slovene language*. Slavica Publishers, Columbus, Ohio.

Runjing Liu, Jeffrey Regier, Nilesh Tripuraneni, Michael Jordan, and Jon Mcauliffe. 2019. Rao-Blackwellized stochastic gradients for discrete distributions. pages 4023–4031.

Horace G. Lunt. 2001. *Old Church Slavonic Grammar*, seventh revised edition. Mouton de Gruyter, Berlin and New York.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9:2579–2605.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. *ICLR*.

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2019. Ab antiquo: Proto-language reconstruction with RNNs. *arXiv preprint arXiv:1908.02477*.

Aanatolij K. Mojsijenko et al. 2010. *Sučasna ukraïns'ka literaturna mova : pidručnyk*. Znannja, Kyïv.

Yugo Murawaki. 2017. Diachrony-aware induction of binary latent representations from typological features. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 451–461.

Yugo Murawaki and Kenji Yamauchi. 2018. A statistical model for the joint inference of vertical stability and horizontal diffusibility of typological features. *Journal of Language Evolution*, 3(1):13–25.

Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649.

Emmanuel Paradis, Simon Blomberg, Ben Bolker, Joseph Brown, Julien Claude, Hoa Sien Cuong, and Richard Desper. 2019. Package 'ape'. *Analyses of phylogenetics and evolution, version*, 2(4).

Maks Pleteršnik. 1894. *Slovensko-nemški slovar*. Knezoškofijstvo, Ljubljana.

Simone Pompei, Vittorio Loreto, and Francesca Tria. 2011. On the accuracy of language trees. *PloS one*, 6(6).

Bernd Pompino-Marschall, Elena Steriopolo, and Marzena Żygis. 2017. Ukrainian. *Journal of the International Phonetic Association*, 47(3):349–357.

Naruya Saitou and Masatoshi Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425.

Ernest A. Scatton. 1984. *A reference grammar of modern Bulgarian*. Slavica Publishers, Columbus, Ohio.

Heinz Schuster-Šewc. 1968. *Gramatika hornjoserbskeje rěče : 1 : Fonematika a morfologija*. Nakładnistwo Domowina, Budyšin.

George Y. Shevelov. 1964. *A prehistory of Slavic : the historical phonology of common Slavic*. Winter, Heidelberg.

Rachit Singh, Jeffrey Ling, and Finale Doshi-Velez. 2017. Structured variational autoencoders for the beta-bernoulli process. In *31st Conference on Neural Information Processing Systems*.

Martin R. Smith. 2019. Bayesian and parsimony approaches reconstruct informative trees from simulated morphological datasets. *Biology Letters*, 15(2):20180632.

Elena Stadnik-Holzer. 2009. Artikulatorische Phonetik. In Sebastian Kempgen, Peter Kosta, Tilman Berger, Karl Gutschmidt, and Sebastian Kempgen, editors, *Die slavischen Sprachen / The Slavic Languages*. De Gruyter Mouton, Berlin, Boston.

Anant Subramanian, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard Hovy. 2018. SPINE: Sparse interpretable neural embeddings. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Elmar Ternes and Tatjana Vladimirova-Buhtz. 1990. Bulgarian. *Journal of the International Phonetic Association*, 20(1):45–47.

Jörg Tiedemann. 2018. Emerging language spaces learned from massively multilingual corpora. In Eetu Mäkelä, Mikko Tolonen, and Jouni Tuominen, editors, *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018)*, pages 188–197.

Max Vasmer. 1953-1958. *Russisches etymologisches Wörterbuch*, volume 1-3. Winter, Heidelberg.

Rastislav Šuštaršič, Smiljana Komar, and Bojan Petek. 1995. Slovene. *Journal of the International Phonetic Association*, 25(2):86–90.

Søren Wichmann, Eric W. Holman, Taraka Rama, and Robert S. Walker. 2011. Correlates of reticulation in linguistic phylogenies. *Language Dynamics and Change*, 1:205–240.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Shijie Wu and Ryan Cotterell. 2019. Exact hard monotonic attention for character-level transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1530–1537, Florence, Italy. Association for Computational Linguistics.

Irena Yanushevskaya and Daniel Bunčić. 2015. Russian. *Journal of the International Phonetic Association*, 45(2):221–228.